A standardized and reproducible method to measure decision-making in mice

The International Brain Laboratory[1], Valeria Aguillon-Rodriguez, Dora E. Angelaki, Hannah M. Bayer, Niccolò Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle A. Chapuis, Anne K. Churchland, Yang Dan, Eric E. Dewitt, Mayo Faulkner, Hamish Forrest, Laura M. Haetzel, Michael Hausser, Sonja B. Hofer, Fei Hu, Anup Khanal, Christopher S. Krasniak, Inês Laranjeira, Zachary F. Mainen, Guido T. Meijer, Nathaniel J. Miska, Thomas D. Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Josh I. Sanders, Karolina Z. Socha, Rebecca Terry, Anne E. Urai, Hernando M. Vergara, Miles J. Wells, Christian J. Wilson, Ilana B. Witten, Lauren E. Wool, Anthony Zador

Dept. of Molecular and Cell Biology, University of California, Berkeley, CA, USA
Champalimaud Centre for the Unknown, Lisbon, Portugal
Cold Spring Harbor Laboratory, NY, USA
Zuckerman Institute, Columbia University, NY USA
Center for Neural Science, New York University, NY, USA
Princeton Neuroscience Institute, Princeton University, NJ, USA
Sanworks LLC, NY, USA
Sainsbury-Wellcome Centre for Neural Circuits and Behavior, University College London, London, UK
UCL Institute of Ophthalmology, University College London, London, UK
UCL Queen Square Institute of Neurology, University College London, London, UK

Correspondence: info+behavior@internationalbrainlab.org

Abstract

Progress in neuroscience is hindered by poor reproducibility of mouse behavior. Here we show that in a visual decision making task, reproducibility can be achieved by automating the training protocol and by standardizing experimental hardware, software, and procedures. We trained 101 mice in this task across seven laboratories at six different research institutions in three countries, and obtained 3 million mouse choices. In trained mice, variability in behavior between labs was indistinguishable from variability within labs. Psychometric curves showed no significant differences in visual threshold, bias, or lapse rates across labs. Moreover, mice across laboratories adopted similar strategies when stimulus location had asymmetrical probability that changed over time. We provide detailed instructions and open-source tools to set up and implement our method in other laboratories. These results establish a new standard for reproducibility of rodent behavior and provide accessible tools for the study of decision making in mice.

[1] Authors are listed alphabetically.

Introduction

Reproducibility is a concern across science [Ioannidis, 2005] and has been particularly elusive in the measurement of mouse behavior. Even seemingly simple behavioral responses to pain or stress can be swayed by uncontrolled factors such as the identity [Chesler et al., 2002] or sex [Sorge et al., 2014] of the experimenter. Behavioral assays have also been frustratingly hard to reproduce across laboratories [Chesler et al., 2002; Tuttle et al., 2018] even when they share a similar apparatus [Crabbe et al., 1999]. To study the neural basis of behavior, we need paradigms that reliably elicit and predictably manipulate mouse behavior. The absence of these paradigms has thus limited the use of mouse models in neuroscience. With the growing arsenal of genetic, imaging, and physiological tools available to study mouse brains, developing reliable and reproducible paradigms for studying mouse decision making has only become more urgent.

The International Brain Laboratory (IBL) [International Brain Laboratory, 2017] faced this challenge directly upon inception. A founding goal of IBL is to reveal the neural basis of decision-making by exploring the same mouse behavior across experimental laboratories. This required the IBL to design and deploy a decision-making task in mice that could be reproduced across all of our laboratories. The task should be simple enough for mice to learn quickly, but also intricate enough to expose the neural computations that support perceptual decision-making, and easily extended to study further aspects of perception and cognition. The task should place specific sensory and motor demands on the mouse in each trial over hundreds of trials, ideally providing stronger behavioral control than the assays used in previous attempts to identify sources of behavioral variability [Chesler et al., 2002; Tuttle et al., 2018; Crabbe et al., 1999]. Based on these criteria, the IBL adopted a task in which mice detect the presence of a visual grating of variable contrast in their left or right visual field, and move the grating via a steering wheel to the center of the visual field [Burgess et al., 2017]. This paper describes two implementations of the task: a *basic* version, in which the probability of stimulus appearance at each of two locations is symmetric, and a *biased* version of the task, in which probability of stimulus appearance at the two locations is asymmetric and changes across blocks of trials.

Here we present results from a cohort of mice trained in the basic and the biased versions of the task, demonstrating reproducible behavior across laboratories. To facilitate reproducibility, we specified all rig hardware and software components, we standardized surgical and animal habituation protocols, and we developed an automated pipeline for advancing animals through training. In parallel, we built a system for storing and sharing data [International Brain Laboratory, 2019]. Not only did mice learn the basic version of the task in all laboratories, but critically, they showed asymptotic performance that was indistinguishable across laboratories. Moreover, mice in different laboratories adopted similar strategies when confronted with the biased version of the task. These results indicate that a decentralized and automated experimental pipeline can yield reproducible mouse behavior across laboratories making it now possible to perform high quality, large scale studies of how brain activity gives rise to behavior.

Results

We trained 101 C57BL/6J mice across 7 laboratories in a behavioral task that requires detection of a static visual grating of varying contrast in either the left or right visual field (**Figure 1a**). The visual stimulus is coupled with movements of a response wheel, and animals indicate their choices by turning the wheel left or right to bring the grating to the center of the screen [Burgess et al., 2017]. The visual stimulus appears on the screen after an auditory "go cue" indicating starting the trial and only if the animal holds the wheel for 0.2-0.5 sec. Correct decisions are rewarded with sweetened water (10% sucrose solution) [Guo et al, 2014], while incorrect decisions are indicated by a noise burst and are followed by a longer inter-trial interval (2 s). We first present results obtained in the *basic* version of the task, where the probability of a stimulus appearing on the left or the right is 50:50. We then present preliminary results obtained in the *biased* version of the task, a variant where probability switches in blocks of trials between 20:80 favoring the right and 80:20 favoring the left.

Standardized behavioral apparatus and behavioral training

To facilitate the reproduction of behavioral performance across laboratories, we standardized variables that we thought would be critical and that we could readily control, such as animal strain and provider, age range, weight range, light-dark cycle, water access, food protein and fat. Other variables of interest were harder to control, but were nonetheless measured (e.g., temperature, humidity, environmental sound, etc., **Suppl. Table 1**).

Every animal was subject to standardized procedures and was trained in a standardized setup (**Figure 1b,c**). First, we performed surgery to implant a headbar for fixation, following a standardized surgery protocol (**Appendix 1**). Then, during the subsequent recovery period we handled the mice and weighed them daily. Following recovery, we put mice on water scheduling and habituated them to the experimental setup, following standardized procedures (**Appendix 2**). The experimental setups were identical across laboratories, built from open-source hardware and software with identical components (**Appendix 3**). These components include systems for head-fixation, visual and auditory stimuli presentation, video and audio recording, and measurement of ambient temperature and humidity.

The training proceeded in automated steps, following predefined criteria (**Figure 1d-f**). Initially, mice experienced only "easy" trials with highly visible stimuli (100% and 50% contrast). When animals met predefined performance criteria based on the proportion of correct responses over the recent trial history (details in **Appendix 2**), the contrast set was incrementally updated to include contrasts of 25%, 12%, 6%, and finally 0% (**Figure 1d, Suppl. Table 2**). For the example animal in Figure 1, the 25% contrast trials were introduced in session 10, 12% in session 12, and the remaining contrasts in session 13 (at which point we removed the 50% contrast trials, because well-trained mice performed as well at 50% as they did at 100% contrast). To reduce response biases, incorrect responses on "easy" trials were followed by a "repeat trial" where the same stimulus location was more likely to be repeated.

As training progressed, the automated protocol gradually increased motor demands and reduced rewards, to encourage determined responses and increase the number of trials. At the beginning of training, to encourage the mice, the wheel gain was high (8 deg/mm), making the stimuli highly

responsive to small wheel movements, and correct responses were rewarded with large volumes (3 μL). As training progressed, the wheel gain was reduced to 4 deg/mm and reward volume to 1.5 μL according to a predefined schedule.

The duration of training sessions was not fixed, but varied according to performance. Sessions lasted at most 90 minutes and ended according to various criteria based on number of trials, total duration, and response times (**Suppl. Table 3**). For instance the example animal in Figure 1, session 2 reached criterion when 45 minutes elapsed with <400 trials; sessions 7, 10, and 14 ended when response duration increased to five times above baseline (**Figure 1e**).

Over the course of training, mouse performance gradually improved until meeting final criteria (**Figure 1f**). At first, performance hovered around or below 50%. It could start below 50% because of no-response trials, of systematic response biases, and of our bias-correcting algorithm that favored repeating "easy trials" on the side that the animal tended not to choose. Performance then typically increased so that mice made only rare mistakes (lapses) on "easy trials", as shown for the example mouse on day 14. Animals were considered trained on the basic task when they had been introduced to all contrast levels and had met or exceeded the pre-defined criteria for sustained performance levels, bias, threshold and lapse frequency, for 3 consecutive sessions.

This automated training procedure is described in detail in **Appendix 2**.

Animal performance was monitored daily, entered into a joint database [International Brain Laboratory, 2019] and reviewed in weekly meetings attended by researchers across the collaboration. Metadata about each session and mouse (e.g., session start time, animal weight, etc.) were entered into a colony management database (Alyx) accessible through a web browser [International Brain Laboratory, 2019]. The raw behavioral responses and compressed videos and audio were transferred to a central repository. Behavioral data were automatically ingested into the DataJoint platform [Yatsenko et al, 2018] where automated analyses produced daily visualizations that are now freely accessible online (data.internationalbrainlab.org). This daily analysis of the data allowed us to assess whether animals met predefined criteria for learning and could be considered fully trained in the basic task.

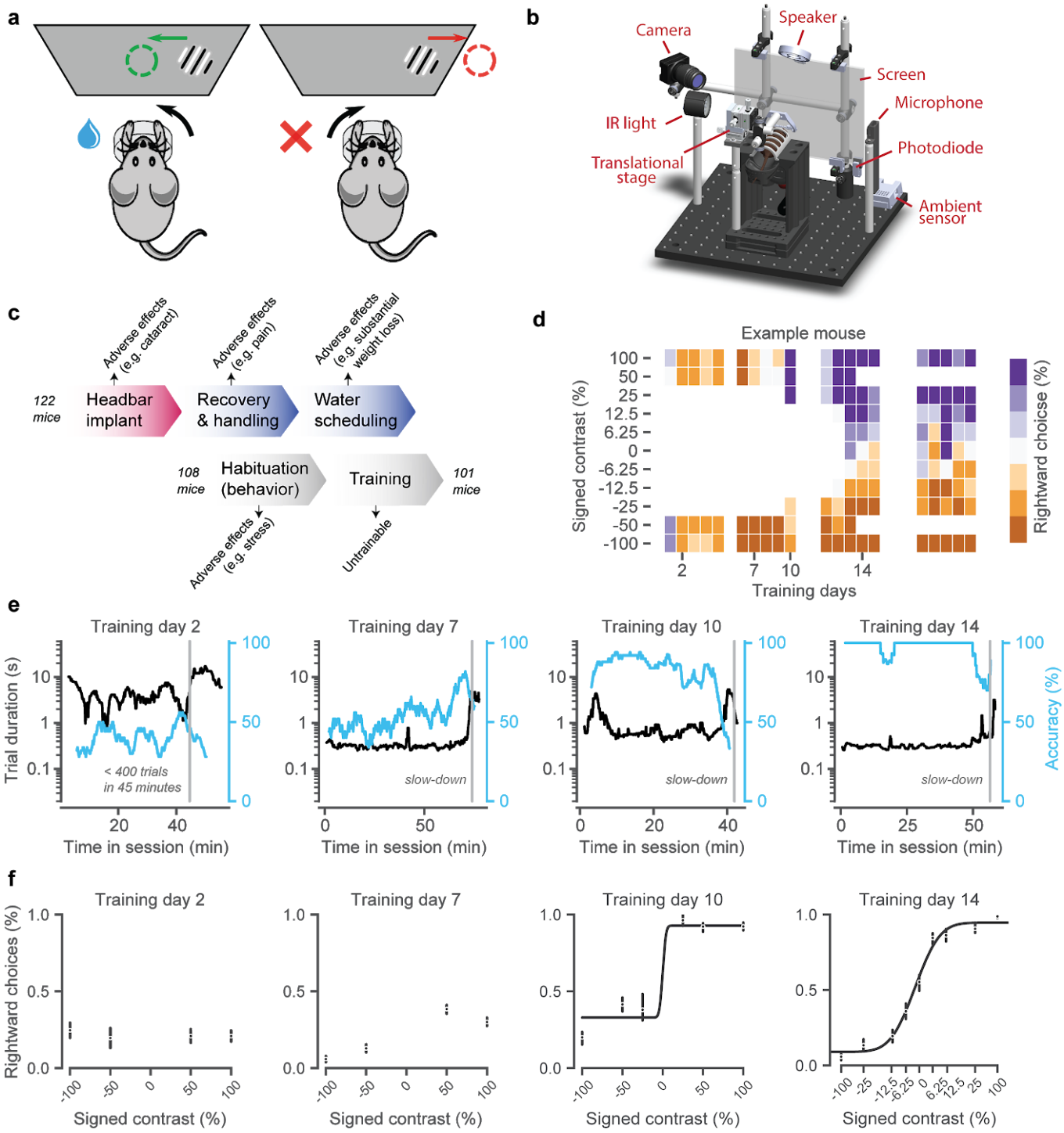


Figure 1. Standardized behavioral apparatus and automated training. a, Schematic of the task. Moving the wheel to bring the visual stimulus to the center of the screen leads to a reward (left panel), whilst bringing it outside leads to a time out (right panel). **b**, CAD model of the behavioral apparatus. **c**, Pipeline for mouse surgeries and training. Of 122 implanted mice, 108 started the training procedure, and 101 completed the training procedure and met the criteria for training in the basic task. **d**, Performance of an example mouse (KS014, from Lab 1) throughout training. Squares indicate choice performance for a given

stimulus on a given day. The color scale indicates the percentage of right (purple) and left (brown) choices. Empty squares indicate stimuli that were not presented. Negative contrasts denote stimuli on the left, positive contrasts denote stimuli on the right. **e,** Example sessions from the same mouse. *Vertical lines* indicate when the mouse reached the session-ending criteria based on response time (*black*) and response accuracy (*cyan*) averaged over a rolling window of 10 trials. **f**, Psychometric curves for those sessions, showing fraction of rightward choice as a function of stimulus position and contrast. Circles show the mean and error bars show ± 1 S.D.

Learning rates differ across mice and laboratories

Nearly all the mice that started the training process reached the criteria for being trained (101 out of 108). However, the learning rates of these trained mice were highly variable both within and across laboratories (**Figure 2a, b**). This was perhaps to be expected, as we did not prescribe any learning rate criteria: our aim was rather to maximize the number of trained animals. For mice that learned the task, the average training took 18.1 ± 13.5 days (s.d., n = 101), similar to the 14 days of the example mouse from Lab 1 (**Figure 2a**, *black*). The fastest learners met training criteria in 3 days, the slowest 59 days. Most mice that ended up learning the task did so within 40 days (**Figure 2c**). The number of days needed to meet training criteria were significantly different across laboratories (**Figure 2d**, $p = <0.001$, Kruskal-Wallis nonparametric test followed by a post-hoc Dunn's multiple comparisons test, **Suppl. Table 4**). Some labs had fairly homogeneous training rates (e.g., Lab 2 within-lab interquartile range of 7.5 days for animals to reach trained criteria), while other labs had larger variability (e.g., Lab 6 interquartile range of 22 days). Variability in performance was large in the middle of training but decreased as animals learned the task. For example, the variance (s.d.) of performance on easy trials was 19.1% (mean 80.7%) on day 15, but 10.1% (mean 91.1%) on day 40 (**Suppl. Figure 4a-b**). Some variability in performance, however, persisted after training, with the performance of some trained animals occasionally dropping. For instance, this occurred in multiple mice in Lab 4 (**Figure 2a**). Such drops could be due to occasional losses of motivation resulting in no-response trials, which are labeled as incorrect trials in our analyses.

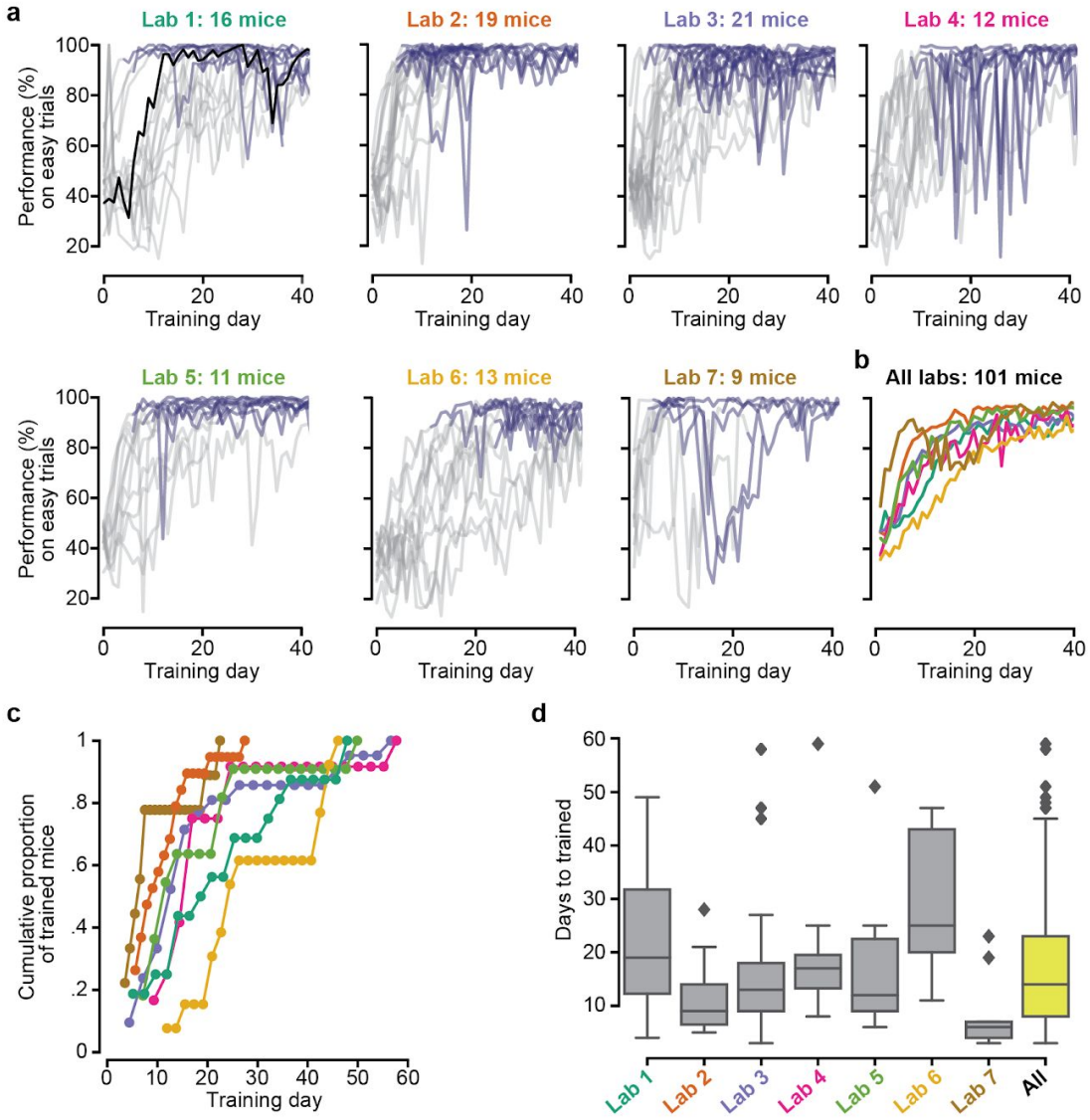


Figure 2. Learning rates differ across mice and laboratories. a, Performance on easy contrast trials (50% and 100% contrast) across mice and laboratories. Each panel represents a different lab, and each curve represents a mouse (*gray*). The transition from *gray* to *blue* indicates when performance criteria for the basic task are met. *Black*, performance for example mouse in Fig. 1d-f. **b,** Average performance for each laboratory across training days. **c,** Cumulative proportion of mice to have reached trained status as a function of session number. **d**, Distribution of training times by laboratories (*gray*) compared to the IBL as whole (*yellow*). Outliers are presented as diamonds.

Performance of trained mice is indistinguishable across laboratories

Despite the variability in learning rates, once animals were trained their performance on completed trials was indistinguishable across laboratories (**Figure 3a-e**). In every laboratory, psychometric curves showed a stereotyped shape (**Figure 3a**). The averages across mice of these psychometric curves were similar across laboratories (**Figure 3b**). The asymptotic values of these curves (i.e., the responses on easy contrasts of 50% and 100%) were close to perfect (90.3 ± 3.6% correct, **Figure 3c**) with no significant difference across the 7 laboratories. The slope of the curves, which measures contrast

sensitivity, was also similar across laboratories, at 13.6 ± 4 (s.d., n = 7, **Figure 3d**). Finally, the horizontal displacement of the curve, which measures response bias, was small at 0.1 ± 7.8 (s.d., n = 7, **Figure 3e**). All these measures showed no significant difference across laboratories ($p > 0.05$).

Variations across laboratories were also small in terms of trial duration and number of trials per session, even though no specific effort was made to harmonize these variables (**Suppl. Figure 2a,b**). The median trial duration was 427 ± 242 ms, showing small differences across laboratories (**Figure 3f**, $p = 0.006$, Kruskal-Wallis nonparametric test **Suppl. Table 5**). Mice on average performed 725 ± 240 trials per session, with occasional differences across laboratories (**Suppl. Figure 2b**). This difference was significant ($p = 0.001$, one-way ANOVA) but only in respect to one laboratory relative to the rest (**Suppl. Table 6**). It may reflect true differences between the mice in the laboratories, but may also simply reflect different experimenter decisions on when to end training sessions: our standard protocol suggested but did not mandate when to end a session.

Having found little variation in behavioral variables when considered one by one, we next asked whether in combination they may define a pattern that would distinguish laboratories from each other. Considering the three behavioral variables in **Figure 3c-e**, we trained a classifier (Random Forests, [Pedregosa et al., 2011]) to predict lab membership from these behavioral variables. To ensure robustness, we used 3-fold cross-validation with 2,000 splits of test and training sets, which resulted in a large distribution of classifier results. Chance level was determined by shuffling the laboratory labels and repeating the classification as before.

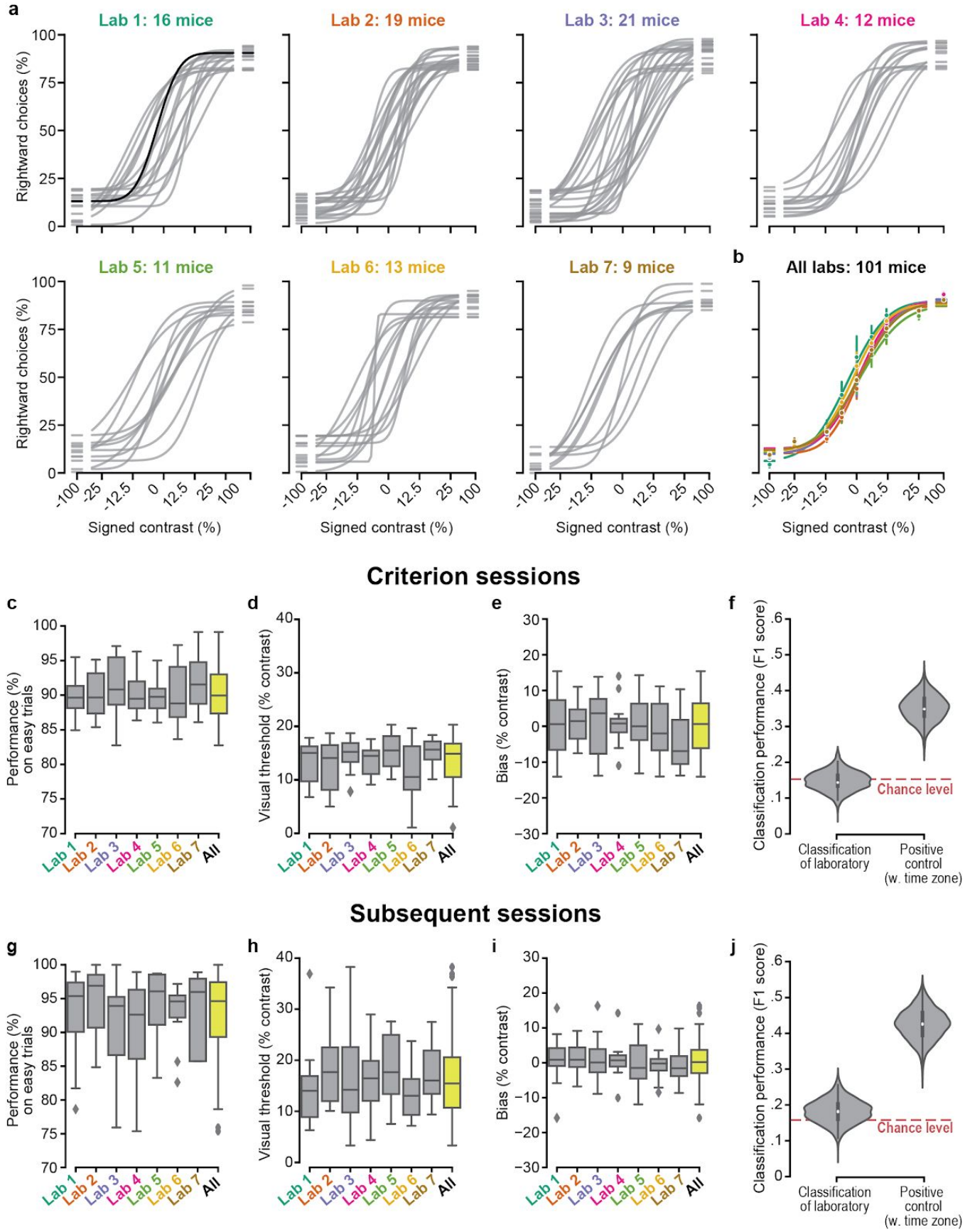

a
Lab 1: 16 mice
Lab 2: 19 mice
Lab 3: 21 mice
Lab 4: 12 mice
Lab 5: 11 mice
Lab 6: 13 mice
Lab 7: 9 mice
b
All labs: 101 mice
Rightward choices (%)
Signed contrast (%)
Criterion sessions
c
Performance (%) on easy trials
d
Visual threshold (% contrast)
e
Bias (% contrast)
f
Classification performance (F1 score)
Chance level
Classification of laboratory
Positive control (w. time zone)
Subsequent sessions
g
h
i
j
Lab 1
Lab 2
Lab 3
Lab 4
Lab 5
Lab 6
Lab 7
All

Figure 3. Indistinguishable psychophysical performance in trained mice across laboratories. a, Psychometric curves across mice and laboratories in trained mice. Each line represents an individual mouse (grey). Black line shows psychometric curve for the example mouse in Fig. 1d. **b**, Average psychometric curve for each laboratory. Circles show the mean and error bars ± 1 the S.D. **c-e**, For the three sessions at which a mouse passes the "trained" criterion, box plots showing the distribution of (**c**) performance on easy trials (50 and 100% contrasts), (**d**) visual threshold, and (**e**) bias across laboratories (grey) and IBL as a whole (yellow). **h**, Performance of a Random Forest classifier when trying to predict in which lab mice were trained, based on the behavioral metrics in **c**-**e**. We included the timezone of the laboratory as a positive control. Dashed red line represents chance-level classification performance. **g-i**, For the subsequent sessions after criterion (14 ± 3.8 sessions per animal), using only data from the unbiased block (first 90 trials per session), box plots as in **c-e**. **j**, Performance of a Random Forest classifier trying to predict lab membership based on the behavioral metrics in **g-i**.

The classifier failed to identify the laboratory of origin of the mice, indicating that there were no systematic behavioral signatures of mice in one laboratory over others (**Figure 3c-e**). The classifier performed at chance level (**Figure 3f**, *left*), with an F1 score of 0.14 ± 0.027 and a 95th percentile of the distribution included the chance level of 0.15 (estimated by shuffling the laboratory labels). As depicted by the confusion matrix, the most common classifications were off-diagonal, and hence incorrect (**Suppl. Figure 3b**). For instance, mice from Lab 6 were often incorrectly classified as being from Lab 2. As a positive control, we included an informative variable: the time zone in which animals were trained. In this control, the classifier performed well above chance (**Figure 3f,** *right*). Similar results were obtained with two other classifier algorithms (**Suppl. Figure 3**).

For many applications, it is important to know that the performance of trained mice is not only reproducible, but also stable in once the animals are trained. To assess this stability, we compared the behavioral performance across laboratories 1-15 days after reaching trained status (**Figure 1c, Figure 3g-i**). Again, we observed that lab membership could not be identified above chance by classifiers (**Figure 3j, Suppl. Fig 3d-f**) and there was no significant difference across laboratories on the performance, visual threshold, and bias (**Figure 3g-i**).

Mice successfully integrate priors into their decisions and task strategy

In addition to the basic task, we implemented a biased variant of the task, an extension that includes more complex across-trial dynamics (**Figure 4a**,**b**). This task variant enabled us to examine how animals integrate information across trials, and use this prior knowledge in their perceptual decisions. Once animals reached satisfactory performance in the basic task, as determined by our predetermined criteria (**Suppl. Table 2**), we introduced a biased prior over stimulus location. Sessions started with a block of unbiased trials (50:50 probability of left vs. right, **Figure 4a**, *gray region*) and then alternated blocks of variable length (20-100 trials) biased towards the right (20:80 probability, **Figure 4a**, *red regions*) or towards the left (80:20 probability, **Figure 4a**, *blue regions*). The transition between blocks was not signaled, so the mice had to estimate a prior for stimulus location based on recent task statistics. This task probes if and how mice learned to make decisions that combine incoming visual evidence with internal beliefs about the dynamic structure of the environment. To assess performance we thus measured the change in rightward choices as a function of stimulus contrast and block type (80:20 vs. 20:80; **Figure 4b,** *distance between black dotted lines*).

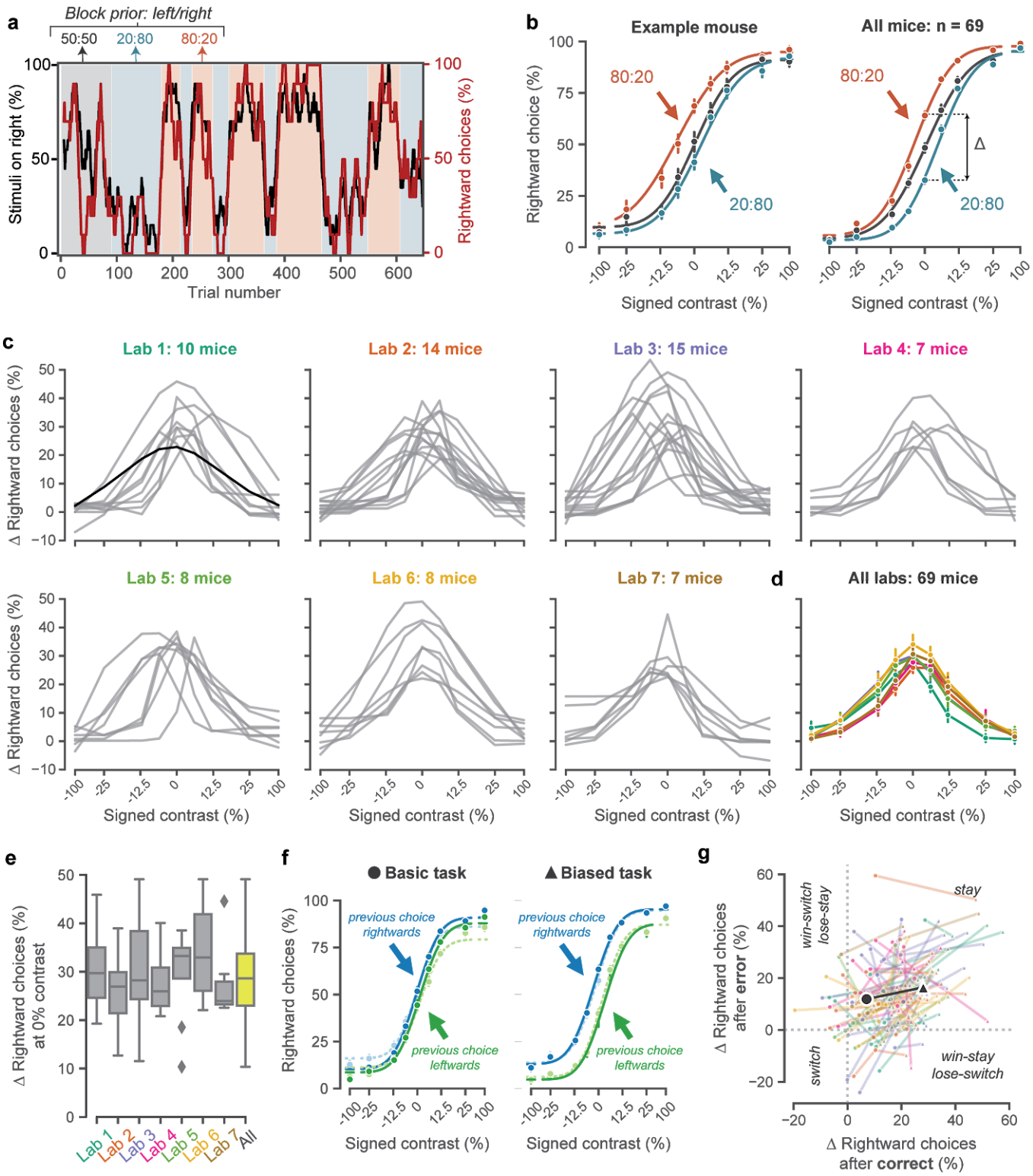


Figure 4. Mice successfully integrate priors into their decisions and task strategy. a, Block structure in an example session. Each session started with 90 trials of 50:50 prior probability, followed by alternating 80:20 and 20:80 blocks of varying length. Presented stimuli (black, 10-trial running average) and the animal's choices (red, 10-trial running average) track the block structure. **b**, Psychometric curves shift between biased blocks for the example mouse (left) and averaged over all animals (right). For each animal and signed contrast, we computed their 'bias shift' by reading out the difference in choice fraction

between the 80:20 and 20:80 blocks (dashed lines). **c**, Shift in rightward choices as a function of signed contrast. Each line represents an individual mouse (grey), with the example mouse in black. **d**, Average shift in rightward choices as a function of signed contrast for each laboratory (colors as in c; error bars show mean +- 68% CI). **e**, Shift in rightward choices at signed contrast = 0, for each lab (grey) and the IBL as a whole (yellow). **f**, Mice change their strategy across tasks. Psychometric curves as a function of choice and reward history, with previous rightward/leftward choices shown in blue/green, and previous rewarded/unrewarded trials in dark/light colors and solid/dashed lines. Error bars show mean +- 68% CI across animals. **g**, Each animal's 'history strategy', quantified as the shift in the psychometric function by previous choice, separately after correct and error trials and between the basic (circles) and biased (triangles) task; colors as in c.

The change in block statistics caused comparable shifts in the psychometric curves of mice in all 7 laboratories (**Figure 4c,d**). As expected, block structure had the greatest impact on choices when sensory evidence was absent (contrast = 0%, **Figure 4c,d**). In this condition, rightward choices in the two conditions differed by an average of 29.6%, and this value did not significantly differ across laboratories (**Figure 4e**, one-way ANOVA $F(6) = 1.179$, $p = 0.333$).

Lastly, we compared how stimulus statistics influenced choice behavior in both of the task variants (**Figure 4f,g**). For both the unbiased and biased versions of the task, we constructed psychometric functions conditioned on the previous trial's choice (**Figure 4f**, *blue* vs *green*) and outcome (**Figure 4f**, *dark* vs *light*). Mice tended to repeat their choices, both after rewarded and unrewarded choices. The tendency for repetition was weak in the basic task (**Figure 4f**, *left*) and much stronger in the biased task (**Figure 4f**, *right*). Representing each animal's change in 'history strategy' from the basic task (*circles*) to the biased task (*triangles*) as a vector (**Figure 4g**), we found no significant difference in the vector's norm (one-way ANOVA $F(1,6) = 0.5523$, $p = 0.7663$) or angle (circular Watson-Williams test, $F(1,6) = 0.5601$, $p = 0.7603$) between labs. Mice in different laboratories, therefore, incorporated history into their task strategy in similar ways.

Discussion

Like other scientific disciplines, neuroscience has been subject to concerns about reproducibility (Baker, 2016). This could be due to insufficient standardization of protocols and equipment across laboratories, or to the choice of behaviors that depend on too many internal and external factors. We developed and employed identical experimental equipment and a standard set of protocols to examine whether mouse decision-making can be reproduced across laboratories. We trained 101 mice in this task across 7 laboratories in 3 countries, and obtained ~3 million mouse choices. Once mice learned the task, their performance was indistinguishable across laboratories. Mice in different laboratories had similar psychophysical performance in a purely sensory version of the task, and adopted similar choice strategies in a more advanced version of the task that required tracking the stimulus prior probability.

The most prominent type of variability we observed was in a factor we had not attempted to control: the learning rates of individual mice, both within and across laboratories. While we cannot ascertain the source of cross-laboratory variability in learning rates, we believe the variability might originate from differences in the expertise and familiarity of different labs with visual neuroscience and mouse behavior. We speculate that as experimenters gain more experience, the differences in learning times will decrease. Indeed, one approach to standardizing learning rates might be to introduce full automation in behavioral training, reducing the need for human intervention [e.g., Scott et al., 2013; Poddar et al.,

2013; Aoki et al 2017]. We anticipate that approaches such as self-head fixation, live-in home cage training systems, and individualized dynamic training methods [Roy et al., 2018] used independently or in combination may reduce variability in learning rates.

Nonetheless, the origin of individual differences in learning times is a key question that we plan to investigate. The full dataset contains data from many mice, but the number of animals per laboratory is relatively small and quite variable, as different experimental sites became active in the collaboration at different times. As these laboratories continue to train the much larger cohort of mice needed for our map of brain activity [International Brain Laboratory, 2017], we expect that the increased statistical power in the larger data set will enable us to more extensively examine the sources of variability in learning rates.

In addition to demonstrating reproducibility, we hope that the resources we have created will be useful to the community (see **Appendices** for a detailed description of all aspects of the behavioral apparatus and its associated automated training protocol). The apparatus designs, hardware and software that we used are entirely open-source and modular, allowing adjustments to accommodate different scientific questions. The data are freely accessible at data.internationalbrainlab.org, and include all 3 million choices made by all the mice during the task.

We aim to support the wider adoption of this experimental apparatus and task by other laboratories. To this end, we have released the documentation, protocols, and code required to implement the task, train mice, and analyze the behavioral data. We hope that these resources catalyze the development of new adaptations and variations of our approach, and accelerate the use of mice in high quality, reproducible studies of neural correlates of decision-making.

Acknowledgments

We thank C. Reddy for help developing animal welfare and surgical procedures; G. Bekheet, F. Carvalho, P. Carriço, R. Barrett and D. Halpin for help with hardware design. AEU is supported by the German National Academy of Sciences Leopoldina. LEW is supported by a Marie Skłodowska-Curie Actions fellowship (no. 795846). FC was supported by an EMBO long term fellowship and an AXA postdoctoral fellowship. HMV was supported by an EMBO long term fellowship. MC holds the GlaxoSmithKline / Fight for Sight Chair in Visual Neuroscience. This work was supported by the Simons Collaboration on the Global Brain and the Wellcome Trust.

References

Aoki, R., Tsubota, T., Goya, Y., Benucci, A. (2017) An automated platform for high-throughput mouse behavior and physiology with voluntary head-fixation. *Nat Commun* 8:1196

Baker, M. (2016). 1500 scientists lift the lid on reproducibility. *Nature* 533, 452-454

Burgess, C.P., Lak, A., Steinmetz, N.A., ZatkaHaas, P., Bai Reddy, C., Jacobs, E.A.K., Linden, J.F., Paton, J.J., Ranson, A., Schroder, S., et al. (2017). High-yield methods for accurate two-alternative visual psychophysics in head-fixed mice. *Cell Rep*. 20, 2513–2524.

Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L., and Mogil, J.S. (2002). Influences of laboratory environment on behavior. *Nat Neurosci* 5, 1101-1102.

Crabbe, J.C., Wahlsten, D., and Dudek, B.C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* 284, 1670-1672.

Guo ZV, Hires SA, Li N, O'Connor DH, Komiyama T, et al. (2014) Procedures for behavioral experiments in head-fixed mice. *PloS ONE* 9: e88678.

International Brain Laboratory, Bonacchi, N., Chapuis, G., Churchland, A., Harris, K.D., Rossant, C., Sasaki, M., Shen, S., Steinmetz, N., Walker, E. Y., Winter, O., Wells, M. (2019). Data architecture and visualization for a large-scale neuroscience collaboration, https://doi.org/10.1101/827873

International Brain Laboratory. (2017). An International Laboratory for Systems and Computational Neuroscience. *Neuron*. 2017;96(6):1213–1218. doi:10.1016/j.neuron.2017.12.013

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Med* 2, e124.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Poddar, R., Kawai, R., and Olveczky, B.P. (2013). A fully automated high-throughput training system for rodents. *PLoS ONE*, 8, e83171

Roy, N.A., Bak J.H., Akrami A., Brody C.D., Pillow J.W. (2018). Efficient inference for time-varying behavior during learning. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Scott, B., B., Constantinople, C.M., Erlich, J. C., Tank, D.W., Brody, C.D. (2015). Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *eLife*;4:e11308

Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J.C., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A.P., Quinn, T., Frasnelli, J., Svensson, C.I., Sternberg, W.F., Mogil, J.S., (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* 11, 629 632.

Tuttle, A.H., Philip, V.M., Chesler, E.J., and Mogil, J.S. (2018). Comparing phenotypic variation between inbred and outbred mice. *Nat Methods* 15, 994-996.

Yatsenko, D., Walker, E. Y. & Tolias, A. S. (2018) Datajoint: a simpler relational data model. Preprint at arXiv https://arxiv.org/pdf/1807.11104.pdf.

Methods

All procedures and experiments were carried out in accordance with the local laws and following approval by the relevant institutions such as the Animal Welfare Ethical Review Body in the UK and the Institutional Animal Care and Use Committee in the US.

Animals

Animals (all C57BL6/J mice obtained from Jackson Laboratory or Charles River) were co-housed whenever possible, with a minimum enrichment of nesting material and a mouse house. Mice were kept in a 12-h light-dark cycle, and fed with food that was 5-6% fat and 18-20% protein. See **Suppl. Table 1** for details on standardization.

Surgery

A detailed account of the surgical methods is in **Appendix 1**. Briefly, mice were anesthetized with isoflurane and head-fixed in a stereotaxic frame. The hair was then removed from their scalp, much of the scalp and underlying periosteum was removed and bregma and lambda were marked. Then the head was positioned such that there was a 0 degree angle between bregma and lambda in all directions. The headbar was then placed in one of three stereotactically defined locations and cemented in place. The exposed skull was then covered with cement and clear UV curing glue, ensuring that the remaining scalp was unable to retract from the implant.

Materials and Apparatus

For a detailed parts lists and installation instructions, see **Appendix 3**. Briefly, all labs installed standardized behavioral rigs inspired by Burgess et al., 2017, consisting of an LCD screen (LP097QX1, LG), a custom 3D-printed mouse holder and head bar fixation clamp to hold a mouse such that its forepaws rest on a steering wheel (86652 & 32019, LEGO). Silicone tubing controlled by a pinch valve (225P011-21, NResearch) was used to deliver water rewards to the mouse. The general structure of the rig was constructed from Thorlabs parts and was placed inside an acoustical cabinet (9U acoustic wall cabinet 600 X 600, Orion). LCD screen refresh times were captured with a Bpod Frame2TTL (Sanworks). Ambient temperature, humidity and barometric air pressure were measured with the Bpod Ambient module (Sanworks), wheel position was monitored with a rotary encoder (05.2400.1122.1024, Kubler) connected to a Bpod Rotary Encoder Module (Sanworks). Video of the mouse was recorded with a USB camera (CM3-U3-13Y3M-CS, Point Grey). A speaker (HPD-40N16PET00-32, Peerless by Tymphany) was used to play task-related sounds, and an ultrasonic microphone (Ultramic UM200K, Dodotronic) was used to record ambient noise from the rig. All task-related data was coordinated by a Bpod State Machine (Sanworks).

Habituation, Training and Experimental Protocol

For a detailed protocol on animal training, see **Appendix 2**. Mice were handled for at least 10 minutes and given water in hand for at least for two consecutive days prior to head fixation. On the second of these days, mice were also allowed to freely explore the rig for 10 minutes. Subsequently, mice were gradually habituated to head fixation over three consecutive days (15-20, 20-40, and 60 minutes, respectively), observing an association between the visual grating and the reward location. On each trial, with the steering wheel locked, mice passively viewed a Gabor stimulus (100% contrast, 0.1 cycles/degree spatial frequency, random phase, vertical orientation) presented on a small screen (size: approx. 246 mm diagonal active display area). The screen was positioned 8 cm in front of the animal and centralized relative to the position of eyes to cover ~102 visual degree azimuth. The stimulus appeared for ~10 s randomly presented at -35° (left), +35° (right), or 0° (center) and the mouse received a reward in the latter case (3ul water with 10% sucrose).

On the fourth day, the steering wheel was unlocked and coupled to the movement of the stimulus. For each trial, the mouse must use the wheel to move the stimulus from its initial location to the center to receive a reward. Initially, the stimulus moves 8° per mm of movement at the wheel surface. If the mouse completes at least 200 trials within a session, the gain is immediately halved and remains at 4°/mm for all future sessions. At the beginning of each trial, the mouse must not move the wheel for a quiescence period of 200-500 ms (randomly drawn from an exponential distribution with a mean of 350 ms). If the wheel moves during this period, the timer is reset. After the quiescence period, the stimulus appears on either the left (-35°) or right (+35°) with a contrast randomly selected from a predefined set (initially, 50% and 100%). Simultaneously, an onset tone (5 kHz sine wave, 10 ms ramp) is played for 100 ms. As soon as the stimulus appears, the mouse has 60 s to move the stimulus. If it correctly moves the stimulus 35° to the center of the screen, it receives a 3 μL reward; if it incorrectly moves the stimulus 35° away from the center (20° visible and the rest off-screen), it receives an error timeout. If the mouse responds incorrectly or fails to reach either threshold within the 60-s window, a noise burst is played for 500ms and the inter-trial interval is set to 2 s. If the response was incorrect and the contrast was 'easy' (≥50%), a ‘repeat’ trial follows, in which the previous stimulus contrast and location is presented with a high probability (see **Appendix 2**).

Mice were classified as having learned the basic visuo-spatial detection task once three criteria were met: (1) 0% and 6% contrasts had been introduced to the contrast set, (2) >200 trials were completed with >80% performance on easy (100% and 50% contrasts) trials in each of the last three sessions, and (3) a four-parameter psychometric curve (bias, lapse left, lapse right, threshold) fitted to performance on all trials from the last three sessions had parameter values of bias < 16, threshold < 19, and lapses < 0.2.

Once an animal was classified as trained on the basic task, it moved to a biased version of the visual detection task. In this variant of the task, the trial structure is identical, except that stimuli are more likely to reappear on the same side for variable blocks of trials, and counterbiasing ‘repeat’ trials are not used. Each session begins with 90 trials in which stimuli are equally likely to appear on the left or right (10 repetitions at each contrast), after which the probability of the stimulus appearing on the left alternates

between 0.8 and 0.2 for a given block. The number of trials in each block is drawn from a truncated exponential (range = 20-100, mean 50, τ = 60)

Classification of laboratory membership

Three different classifiers were used to try to predict in which laboratory a mouse was trained based on behavioral metrics: Random Forest, Naive Bayes and Logistic Regression. We used the scikit-learn implementation available in Python with default configuration settings for the three classifiers. The dataset was split into a training set and a testing set according to 3-fold cross-validation, this random split was repeated 2000 times. For every split, the classification accuracy was calculated as the F1 score (equation 1) which is a standard way of measuring a classifier's accuracy. An F1 score of 0 indicates complete misclassification and a score of 1 indicates perfect classification.

$$F1\ score = \left(\frac{2}{recall^{-1} + precision^{-1}} \right)$$ *Equation 1*

Data and code availability

The data can be accessed in two ways (International Brain Laboratory, 2019): via DataJoint and web browser tools at data.internationalbrainlab.org or via Open Neurophysiology Environment (ONE) through FigShare at https://doi.org/10.6084/m9.figshare.11636748. Python scripts to produce all the figures are available at github.com/int-brain-lab/paper-behavior, and a Jupyter notebook for re-creating Figure 2 can be found at https://jupyterhub.internationalbrainlab.org/.

Appendices

Appendices are available online:

Appendix 1

Appendix 2

Appendix 3

CAD drawings of the components listed in the Appendices and used to build the Behavior Rig are also available online

Supplementary Tables

Suppl. Table 1. Standardization

Category	Variable	Standardized	Standard	Recorded
Animal	Weight	Within a range	18 - 30g at headbar implant	Per session
	Age	Within a range	10-12 weeks at headbar implant	Per session
	Strain	Exactly	C57BL/6J	Once
	Sex	No	Both	Once
	Provider	Two options	Charles River (EU) Jax (US)	Once
Training	Handling	One protocol	Appendix 2	No
	Hardware	Exactly	Appendix 3	No
	Software	Exactly	Appendix 3	Per session
	Fecal count	N/A	N/A	Per session
	Time of day	No	As constant as possible	Per session
Housing	Enrichment	Minimum requirement	At least nesting and house	Once
	Food	Within a range	Protein: 18 - 20%, Fat: 5 - 6.2%	Once
	Light cycle	Two options	12 Hr inverted or non-inverted	Once
	Weekend water	Two options	Citric acid water or measured water	Per session
	Co housing status	No	Co-housing preferred, separate problem mice	Per change
Surgery	Aseptic protocols	One protocol	Appendix 1	No
	Tools/Consumables	Required parts	Appendix 1	No

Suppl. Table 2. Training progression criteria

Adaptive parameter	Initial value
Contrast set	[100, 50]
Reward volume	3 μL
Wheel gain	8 deg/mm

Criterion	Outcome
200 trials completed in a session	Wheel gain decreased 4 deg/mm
> 70% correct	Contrast set = [100, 50, 25]
> 70% correct after above	Contrast set = [100, 50, 25, 12.5]
200 trials after above	Contrast set = [100, 50, 25, 12.5, 6.25]
200 trials after above	Contrast set = [100, 50, 25, 12.5, 6.25,0]
200 trials after above	Contrast set = [100, 25, 12.5, 6.25, 0]
200 trials complete in a session & reward volume >= 1.5 μL	Next session decrease reward by 0.1 μL
For each of the last 3 sessions: >200 trials completed, & >80% correct on 100% contrast & all contrasts introduced & psychometric absolute bias <16 & psychometric threshold <19 & psychometric lapse rates < 0.2	Training on the basic task obtained. Proceed to training on the biased task.

Suppl. Table 3. Within-session disengagement criteria

Criterion	Explanation
Session length > 90 min	Session too long
< 400 trials completed & > 45 min elapsed	Not enough trials
> 400 trials & 20-trial rolling median RT > 5x session median RT	Slow-down

Suppl. Table 4. Comparison of training times across laboratories

	P values Dunn's Multiple Comparisons Test, * p < 0.05, ** p < 0.01, *** p<0.001						
	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
Lab 1							
Lab 2	*0.01						
Lab 3	0.25	0.124					
Lab 4	0.894	0.027	0.361				
Lab 5	0.356	0.181	0.958	0.457			
Lab 6	0.107	***<0.001	**0.005	0.103	*0.019		
Lab 7	**0.003	0.357	0.031	**0.007	0.050	**<0.001	

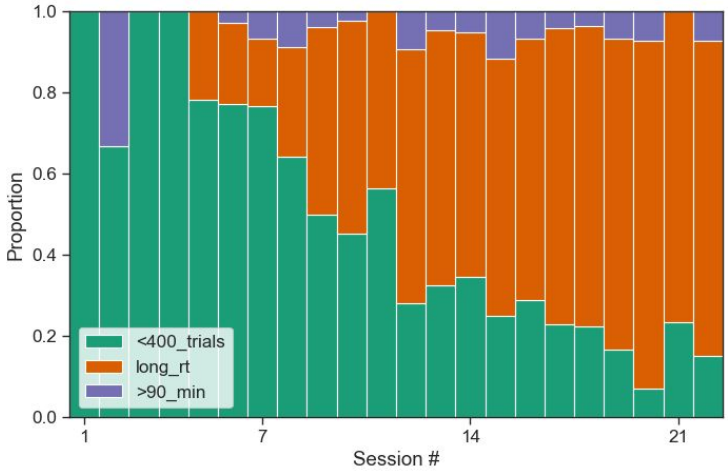
Suppl. Table 5. Comparison of trial completion times across laboratories

	P values Dunn's Multiple Comparisons Test, * p < 0.05, ** p < 0.01, *** p<0.001						
	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
Lab 1							
Lab 2	0.682						
Lab 3	0.590	0.897					
Lab 4	0.062	0.119	0.145				
Lab 5	0.345	0.18	0.143	**0.009			
Lab 6	0.086	0.164	0.197	0.855	*0.014		
Lab 7	**0.004	**0.008	*0.01	0.261	***<0.0001	0.19	

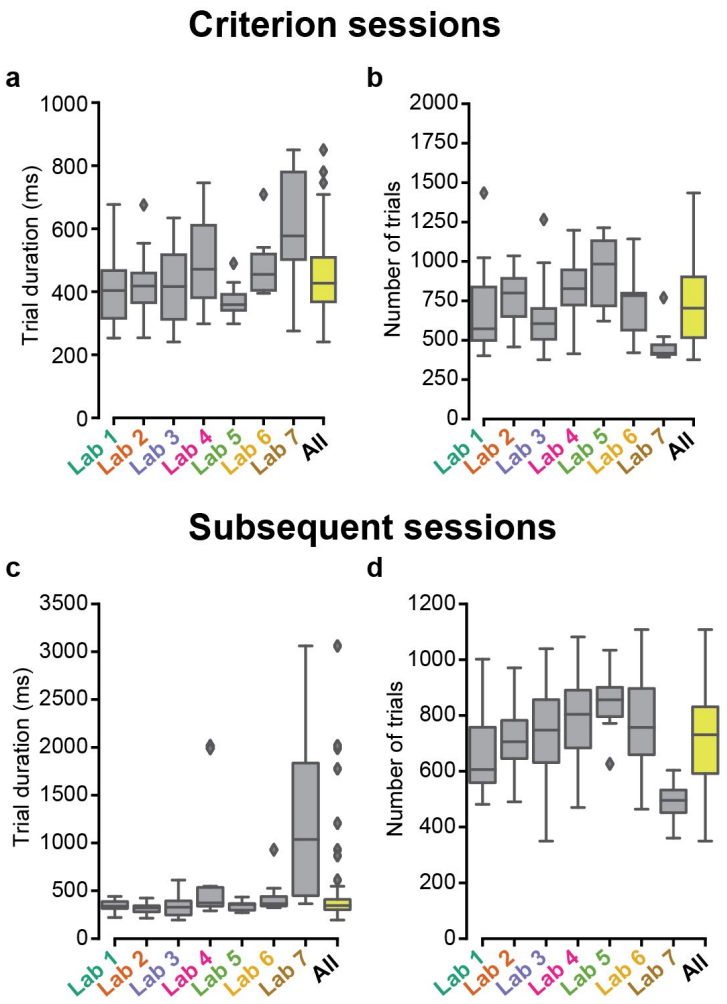
Suppl. Table 6. Comparison of number of trials across laboratories

	P values Tukey's Multiple Comparisons, * p < 0.05, ** p < 0.01, *** p<0.001						
	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
Lab 1							
Lab 2	0.9						
Lab 3	0.9	0.63					
Lab 4	0.686	0.9	0.365				
Lab 5	0.06	0.361	*0.011	0.809			
Lab 6	0.9	0.9	0.9	0.861	0.140		
Lab 7	0.189	*0.016	0.349	*0.007	**0.001	0.147	

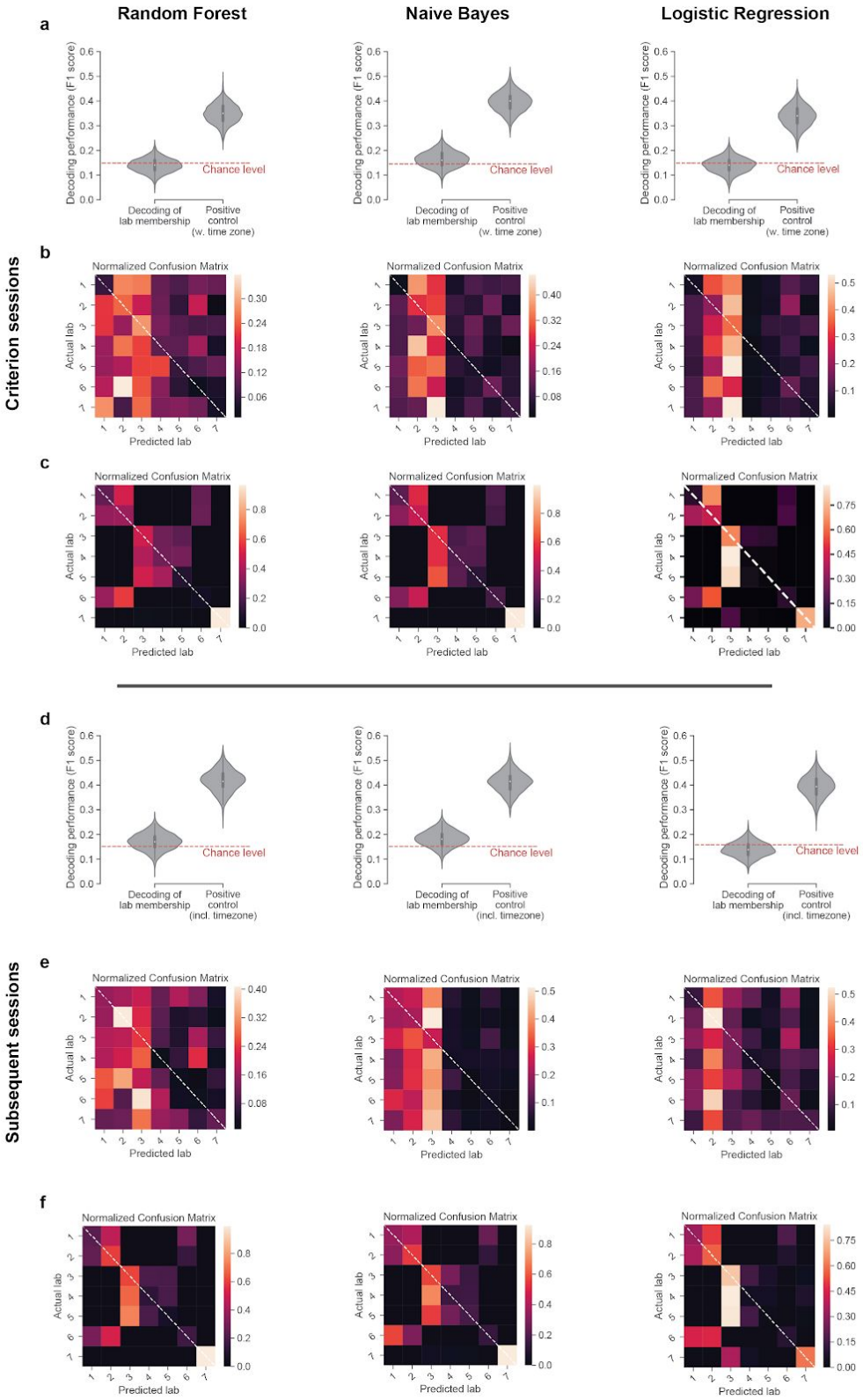
Supplementary Figures



Supplementary Figure 1. Session-ending criteria. Proportion of sessions that ended in each of the 3 criteria for all mice that learned the task. The three criteria were, 1. Fewer than 400 trials in 45 minutes (green); 2. over 400 trials performed and median reaction time over the last 20 trials was over 5x the median for the whole session (orange); 3. Over 400 trials performed and session length over 90 minutes (blue).

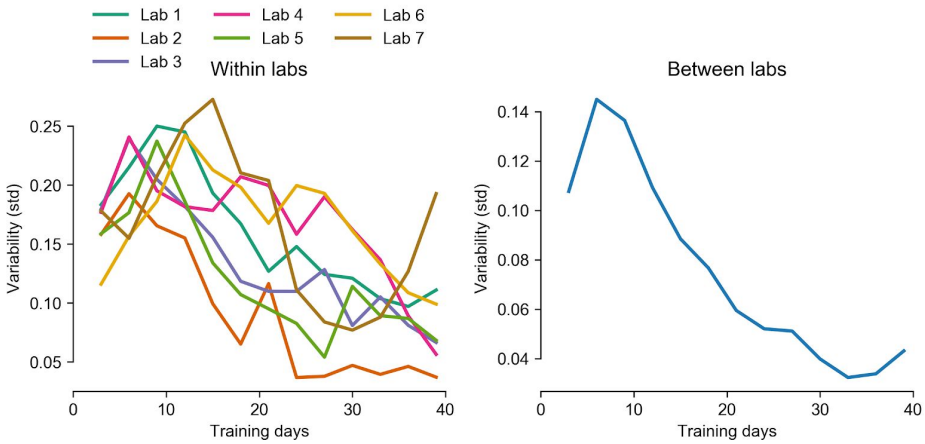


Supplementary Figure 2. Behavioral metrics mice were not explicitly trained on varied over labs. **a-b**, For the three sessions at which a mouse passes the "trained" criterion, box plots showing the distribution of (**a**) trial duration from go cue to correct or incorrect outcome in ms and (**b**) the average number of trials over the three sessions. **c-d**, For the subsequent sessions after criterion (14 ± 3.8 sessions per animal), using only data from the unbiased block (first 90 trials per session). Box plots showing the distribution of (**c**) trial duration as in **a** and (**d**) the average number of trials per session.



Supplementary Figure 3. Classification of lab membership by three different classifiers could not predict lab membership from behavior during criterion or subsequent sessions. a, Cross-validated classification performance of a Random Forest, Naive Bayes and Logistic Regression classifier while predicting lab membership based on behavioral metrics

from Fig. 3c-e (criterion sessions). The positive control included the time zone in which a mouse was trained in the dataset. **b**, Normalized confusion matrices for the classifiers in **a** which indicates the proportion of occurrences that a mouse was classified to be in the predicted lab (x-axis) while it was from the 'actual lab' (y-axis). **c**, Normalized confusion matrices as in **b** for the positive control. **d-e**, Classification performance and confusion matrices as in **a-c** but for the subsequent sessions after criterion had been reached.



Supplementary Figure 4. Performance variability within and across laboratories goes down over training time. a-b, Variability in performance (s.d. of % correct) in easy trials (100% and 50% contrast) (**a**) within, and (**b**) across laboratories during the first 40 days of training.